A Validated Logistic Regression Model to Identify Coronary Heart Disease patients (CHD) within Primary Care Databases in the United Kingdom

K Thiru[1] MPH, P Donnan[2] PHD, F Sullivan[3] MBChB (Hons) PHD
[1] Research Fellow, Fisher Medical Centre Research unit, Skipton, UK. [2] Senior Medical Statistics Lecturer, Tayside Centre of General Practice (TCGP). [3] Prof of General Practice Research and Development, TCGP, University of Dundee, UK.

Abstract

We established the optimal search strategy for identifying coronary heart disease (CHD) patients within the Electronic Patient Record (EPR) of 'paperless' family practices in the UK . Multiple logistic regression modelling (MLRM) and Receiver Operating Characteristic (ROC) curves were used to develop the query. The selected search strategy was validated at 2 additional paperless family practices.

Background

Under NHS directives in the UK, clinicians are increasingly required to practice in a 'paperless' environment and to collect clinical data electronically. These data are required to support budgetary requirements, implementation of Evidence Based Medicine (EBM) and the 'quality cycle'. A primary step to meet these needs is the identification of a target population. Due to the granularity of coding strategies, the structure of the coding frame and the prioritisation and use of READ codes by clinicians, establishing a sensitive (complete) and positive predictive (correct) search strategy is complex. In this context, no technique to establish the optimal search strategy with which to identify a patient population, has been described in the literature. Query strategies have tended to be generic and search for expected diagnostic and prescribing codes. This study attempted to identify and validate the most effective and efficient search strategy for CHD patients in UK family medicine practice.

Method

The model was derived from data from a paperless practice with 13 500 patients. A total of 55 independent variables were identified by a multi-professional primary care team, as desirable for the implementation and management of CHD patients. Of these, 14 redundant or highly correlated codes were removed. Three different ontological representations of CHD were used as the dependent variables (reference standards) for MLRM: 1. a World Heath Organisation (WHO) based definition 2. a national data quality group's definition (PDQ), 3. a model determined by a group responsible for clinical quality of care at the local level (CG). Forward stepwise MLRM using 0.05 and 0.1 entry and exit p-values were used to identify significant codes. The area under the ROC curve (AUC) was used to assess the performance of the models. Age and gender specific models were explored. The final model was for all patients over the age of 35. The number of significant independent variables identified was reduced further using the Receiver Operating Characteristic (ROC) curve. The selected model was validated using data from two independent paperless practices with population s, 11 762 (4-byte) & 7545 (5 byte).

Results

Selection of significant codes was highly influenced by the age and sex of the population considered. Codes selected as optimal for identifying >35 year old patients were dependent on the ontological perspective of the disease used (table 1). PDQ and WHO defined patients were most effectively identified through the select search. The search strategy derived from the PDQ population preformed optimally in relation to sensitivity and yield (1/positive predictive value) statistics on the remaining dependent variables.

Validation site results show the PDQ strategy to perform best when tested on data from a practice using the same Read version. The search was less effective on the newer 5 byte Read version (table 2).

Conclusion

Generic searches may perpetuate inequalities in care i.e. older and female patients may not be identified by non-specific search strategies. MLRM and ROC techniques can be used to identify the optimal search strategies for specific groups of CHD patients. Queries should be bespoke and mindful of the user's conceptualisation of disease. Such techniques can improve practices ability to establish sensitive and high yield positive predictive chronic disease registers. The study explored the wider implications of the results on patient identification and the impact of more granular coding strategies on data collection and queries. When READ codes become incorporated within the SNOMED CT classification system these findings will have international relevance.

Table 1: Summary table of reference standard specific optimal search strategies and summary statistics of effectiveness at FMC (G4 = diagnostic codes for CHD).

Definition	Codes selected	Sensitivity	Specificity	PPV	Yield
WHO	G4, anti arrhythmics and anti-lipids	98.0%	96.4%	71.0%	1.4
PDQ	G4, nitrates and digoxins	99.0%	92.1%	54.1%	1.8
CG	G4, anti hypertensive drugs and hypertension	95.1%	81.0%	35%	2.8

Table 1: Performance of diagnostic code, nitrates and digoxin based queries at training and testing sites

G4 + nitrate + digoxin queries	**WHO**		**PDQ**		**CG2**	
	Sensitivity	PPV	Sensitivity	PPV	Sensitivity	PPV
Training site (4 byte)	99%	51%	99%	54%	86%	52%
Testing 4 byte site	98%	66%	97%	74%	91%	69%
Testing 5 byte site	91%	39%	74%	48%	65%	61%